02450

Introduction to Machine Learning and Data Mining

Project 1

09.03.21



Avi Raj	Piotr Samir Saffrani	Pranjal Garg
s210252	s210364	s210242

Table 1: Students contribution to the report

Section	Avi Raj	Piotr Saffarini	Pranjal Garg
1.Descripiton	40%	20%	40%
2. Detailed explanation	35%	42.5%	22.5%
3. Visualization	15%	42.5%	42.5%
4. Summary	40%	30%	30%

Contents

1 A description of the data set

1.1 Explanation of the data

The South African Heart Disease dataset is about Coronary Heart Disease(hereinafter referred as CHD), basically looking at the population of 462 males in Western Cape, South Africa with a high number of CHD cases. There are roughly 2 control cases per case of CHD in the dataset i.e two non chd cases for one chd case, in order to do a fair statistical analysis and apply machine learning algorithms on the same.

The dataset consists of a variety of attributes theoretically related to CHD from blood pressure measurements to weight metrics(obesity, adiposity) to family history, personality type and age to lifestyle habits (smoking, alcohol consumption). All 10 attributes including whether they have CHD or not, are shown on Table 2.

Table 2: Data features

sbp	systolic blood pressure (mmHg)
tobacco	cumulative lifetime tobacco usage (kg)
ldl	low density lipoprotein cholesterol (mmol/L)
adiposity	The body adiposity is calculated value of body fat (%)
famhist	family history of heart disease (Present, Absent)
typea	type-A behavior
obesity	Body Mass Index (kg/m^2)
alcohol	alcohol consumption (l)
age	age at onset (yrs)
chd	response, coronary heart disease (Positive, Negative)

In some cases the measurements of the above mentioned attributes were taken after the people who had a CHD event (heart attack), had undergone some interventional programs to reduce their blood pressure and other risk factors.

Foremost, the focus of our analysis is to predict a CHD event so that we can help save lives using machine learning. For this we plan to use all the attributes to better predict a person's probability of having a CHD event. Later, we would try to filter down the set of attributes to a smaller subset of attributes which are the key factors in causing CHD. Early prediction of a person's probability of a CHD event would be critical to increase life expectancy of people through disease management.

We would also try to differentiate the attributes in two main categories of it being a reversible attribute or a irreversible attribute and try to predict if they are better markers for predicting a CHD event and thus find out the extent to which people's life can be improved w.r.t to CHD through habit or lifestyle change.

1.2 Reference to origin of the data

We have decided to work with the South African Heart Disease dataset, chosen from the book "The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani and Jerome Friedman(2009) [5].

The dataset provided in the book is a subset of a larger dataset mentioned in the in Rousseauw et al, 1983, South African Medical Journal. The original dataset mentioned in the journal was collected through an intensive postal campaign in 1979 from 82 established target populations predominantly in three Afrikaner communities in the southwestern Cape Province where high occurrences of ischemic heart disease were observed among white Afrikaans-speaking segments of South African society(3,357 white males and 3,8,31 white females). The objective of the survey was to isolate CHD risk factors, mainly focusing on hypercholesterolaemia(ldl), hypertension(sbp) and smoking.

Later, when it was observed that the males for at 2.5 times higher mortality risk than females, Hastie and Tibshirani chose to focus on the male subjects and selected a group of 465 male subjects from the original 3,357 males which later through some probable recording errors was further decreased to 462 subjects (160 case of CHD and 302 controls).

1.3 Previous analysis of the data

The dataset that we have chosen has been analyzed various times and we are going to summarize the paper made by Srisai Sivakumar [6]. The paper tried to use various data mining algorithms to better predict CHD events and looked at various models to figure out the most accurate

model to make predictions.

They started with a pairwise scatter plot of all attributes and then correlation plots to find attributes correlation with one another. Then they did a combined plot of alcohol and tobacco controlled over CHD through colors. They followed this with a combined alcohol-tobacco plot controlled over CHD, but this time they did multiple subplots breaking them down based on age -groups.

Next, they did separate plots for alcohol and tobacco(y axis) against age(x axis), controlled over CHD. Following that, they tried finding influence of family history with CHD and tobacco usage. And in the end, they looked at different classification algorithms like CART, Logistic Regression and Boosting to compare their accuracies.

Some of their key findings were as follows:

- Strong correlation between obesity and adiposity.
- Strong presence of CHD when compared with the combined effect of alcohol and tobacco. compared individually, alcohol doesn't have a strong correlation with the presence of CHD.
- Absence of family history had a higher correlation with the presence of CHD. They attributed it to an indirect effect of higher tobacco consumption in individuals without a family history of CHD.
- Age had an expected effect on the presence of CHD, with low CHD in age groups of 0-20 and 20-30 and increasing effect with an increase in age.
- The difference between the different classification algorithms used was miniscule. Boosting Model gave the best accuracy with 72% followed by Logistic Regression's 70% accuracy and CART and naive baseline model accuracies of around 65%. All of them considered age, tobacco and family history as important predictors.
- The CART model predicted age, tobacco and family history as predictors of CHD. The decision tree, predicting yes in case of absence of family history as explained above.
- The Logistic Regression Model observed that Systolic Blood Pressure (sbp), Alcohol and Adiposity are not significant predictors.
- The Boosting Model deemed systolic blood pressure as a significant predictor at the expense of obesity and adiposity, contrary to the CART and Logistic Regression models.

The study concluded with saying that medical, biological, heredity and lifestyle choices are indeed reliable indicators to predict the presence of heart diseases but with a lack of a clear understanding of how they are classifying- medical, biological, hereditary and lifestyle choices and how exactly did they reach to this conclusion.

1.4 Main objectives to be performed on this data

The main objectives that we want to accomplish from the dataset, like many others before us, is to find better prediction markers for CHD from this dataset. Using machine learning in the biomedical space is a vibrant field with a lot of direct benefits to humanity. So we want to use this project to gain some knowledge about machine learning in the medical field and figure out the challenges and their answers in trying to implement such a solution. Hence we can say our main goal is both educational and health advisory.

Further refining our objectives, we would want to figure out the impact of lifestyle and habits on a person's health specifically their chances of contracting a coronary heart disease as the dataset is limited to heart diseases in this case. We would love to find some new findings that could help doctors and individuals in dealing with coronary heart diseases. We would like to figure out advisories and markers in terms of lifestyle for the younger population so that they can reduce their chances of contracting such a disease which can never be cured but only managed as they grow older. The findings could also serve as a wakeup call for individuals engaged in unhealthy practises.

We plan to approach the problem through a unique set of questions and would try to find answers to these questions. Some of the questions that we might be looking to answer through our analysis are as follows:

- Age wise comparison of CHD for different age groups to find key findings for different age groups separately. Correlation between personality type and CHD and whether reducing competitiveness and the constant need to work would lead to a reduction in the chances, of a person contracting CHD.

- Do a multi attribute correlation analysis for attributes like adiposity, ldl, obesity, blood pressure etc together to see if combining attributes helps us with better predicting CHD.
- Correlation of personality type with other attributes and answer questions like do type A people have a high chances of either using alcohol, tobacco, or being obese and having higher adiposity due to stress etc.
- Predict adiposity(fat percentage) through a regression analysis
- Predict personality type(type a value) through regression analysis models.

The above questions mentioned are just a sample of our larger analysis. In essence, as said earlier, we would like to help people make better lifestyle choices to improve their chances of never contracting a CHD and have a healthier and longer lifespan.

1.5 Regression and classification tasks

In the regression task, we wish to predict adiposity based on the other attributes. We wish to predict adiposity as its a measure of fat percentage of an individual and it is something that can be improved through better food choices and exercise and hence it would be a good thing to know whether changing this one attribute about yourself can help a person improves all their other attributes like blood pressure, chd chances, ldl etc.

In the classification task we wish to predict personality and lifestyles impact on the presence of CHD. The class label we wish to predict is CHD; meaning whether the person has coronary heart disease or not. We are right now looking to make a basic prediction based on all other attributes initially and then further refine it by trying to make a prediction with the most informative subset of attributes and try to do several iterations which would include but is not limited to personality and lifestyle attributes which would mainly comprise of TypeA, adiposity, ldl, obesity and blood pressure and age.

1.6 Necessary data transformations

For the classification task, we would convert the Family History attribute from Present, Absent to a binary number i.e 0 or 1 where 0 would mean an absence of family history for CHD and 1 would mean that the individual has a family history of CHD. Since the class is binary we would not lose any information and doing "one-out-of-K encoding" would just create an unnecessary column without giving much benefits.

For the regression task we would just append the CHD attribute to the attribute table and move over the adiposity attribute to a class label further dividing the adiposity attribute into a class based structure like class 1, class 2, class 3 and so on and where our task would be to predict the adiposity class.

Also since we are looking to do an age based analysis as well we will classify age values to certain classes like adolescence, young adults, adults, middle aged, old etc or we could do it in terms of the decade of life for individuals like 20's, 30's, 40's etc.

2 A detailed explanation of the attributes of the data

2.1 Description of attributes types

Summary of atributes types is shown in the Table 3. In this dataset, sbp takes only whole values so it is discrete. Also, its a ratio because multiplication and divison holds a meaning as it can be said the sbp is twice as much as the other person etc. Morevorer, 0 holds a physical meaning of 0 pressure(mmHg). Features numbered: 2, 3, 8, 9 (tobacco, ldl, alcohol, age) are ratios for the same reason, but only age is discrete, and the rest are continuous, because they can take real values. Adiposity is described in percentages , so it is ratio - it makes sense to compare values by multiplying or dividing. The only Interval feature is obesity. As it is described as BMI, zero has no physical meaning and is arbitrary. Two features which are binary and nominal are family history of CHD (famhist) and if a person has CHD (chd). Both can only take two values: true or false (1 or 0).

2.2 Data issues

The data would be more robust if it was available for a diverse population consisting of people from various demographics, and included women as well. Due to the fact that the samples are all from the same geographic region, environmental factors of CHD could not be controlled. Also, a larger sample size would provide better results of estimation with increased accuracy of the system.

The data set does not contain any missing values and all the data points lie within the range of reason, hence there are no corrupted data points.

2.3 Basic summary statistics of the attributes

Here is a detailed explanation of each attribute of our data and then you can find all the summary statistics in the Table 3:

Sbp - systolic blood pressure - Measures the pressure(in mmHg) of blood when the heart beats. Values below 120 are recognized as normal. Values above 140 signify high blood pressure, causing CHD. Majority of people in this dataset higher sbp than usual, as only 25% of them have it below 124 and half of the people are above 148.

Tobacco - Cumulative consumption of tobacco(in kg) - how many kgs of tobacco a person has smoked in their lifetime. To give perspective, 1.9 kg of tobacco equals approximately consuming 10 cigarettes per day for a year, as 25 grams of tobacco is 50 cigarettes.

Ldl - Low Density Lipoprotein Cholesterol (in mmol/L) - Optimal values of ldl [1] should be below 2.6, and readings above 4.1 and 4.9 are considered high and very high respectively. In this data set, more than half of records have high ldl (4.3 or more).

Adiposity - body adiposity index (%) - Estimates percentage of body mass which is made of fat. Normal values highly depend on the sex and age. Since our dataset contains only men, obesity is considered with adiposity index greater than 29%, overweight more than 23%, healthy man is considered from 11% to 23% [2]. Since our dataset containing data only on men, the normal adiposity should be less than 30%. The values seemed to be evenly distributed in our datase.t

Famhist - Family history of CHD (Positive, negative) - Whether someone within the family of the subject has Coronary Heart Disease or not.

Typea - Type A behaviour personality - The Short Rating Scale is used to determine if person has type A personality. Values can vary from 12 to 84, where anything above 55 is considered a type A behaviour [3].

Obesity - Obesity represented as Body Mass Index(BMI) (kg/m^2) - BMI is defined as the body mass divided by the square of the body height. It has similar meaning to adiposity but doesnt measure body fat percentage. Less than 18.5 is underweight, over 25 is overweight and over 30 is obese. [4].

Alcohol - alcohol consumption (in litres) - based on yearly alcohol consumption in South Africa (9.5 litres/year) [7], We have assumed this feature is cumulative consumption and the mean of consumption equals about two years of consumption, same as tobacco.

Age - age of the person (in years) - The median and mean of age in our data is 45 years old, it is good value for CHD prediction, as the studies show that with age, risk of CHD increases [8]

chd - presence of Coronary Heart Disease (0, 1)- Each man is classified whether they have CHD or not and 35% cases in our dataser have CHD(value 1). For all the 9 attributes the descriptive statistical values are as follows:

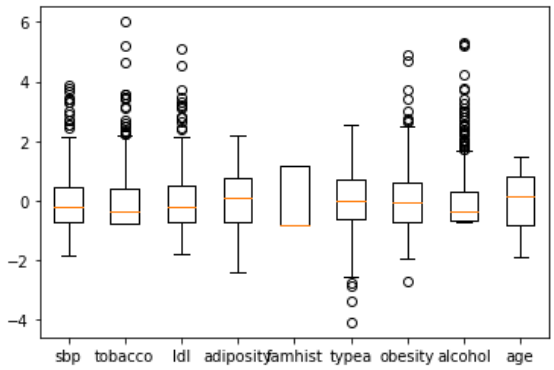
Table 3: basic data description

No.	Attribute Name	Attribute Type	<Min, Max >	Mean	Q1 Median Q3	Standard Deviation
1.	sbp	Discrete, Ratio	<101, 218>	138.32	124 134 148	20.4964
2.	tobacco	Continuous, Ratio	<0, 31.2>	3.64	0.05 2 5.5	4.593
3.	ldl	Continuous, Ratio	<0.98, 15.33>	4.74	3.28 4.34 5.79	2.0709
4.	adiposity	Continuous, Ratio	<6.74, 42.49>	25.41	26.12	7.7806
5.	famhist	Binary, Nominal	<0, 1>	0.42	0 0 1	0.49
6.	typea	Continuous, Ordinal	<13, 78>	53.1	47 53 60	9.8175
7.	obesity	Continuous, Interval	<14.7, 46.58>	26.04	22.99 25.8 28.49	4.2137
8.	alcohol	Continuous, Ratio	<0, 147.19>	17.04	0.51 7.51 23.89	24.4811
9.	age	Discrete, Ratio	<15, 64>	42.8	31 45 55	14.609
10.	chd	Binary, Nominal,	<0, 1>	0.346	0 0 1	0.4763

3 Data Visualization & PCA

3.1 Issues with outliers in the data

Figure 1: Boxplot



Considering a normal distribution, 99.72% of the entire lies within ±3 standard deviations from the mean.Any data beyond this considered as outliers. This value can be easily calculated using the 1st quartile(Q1) and 3rd quartile(Q3).

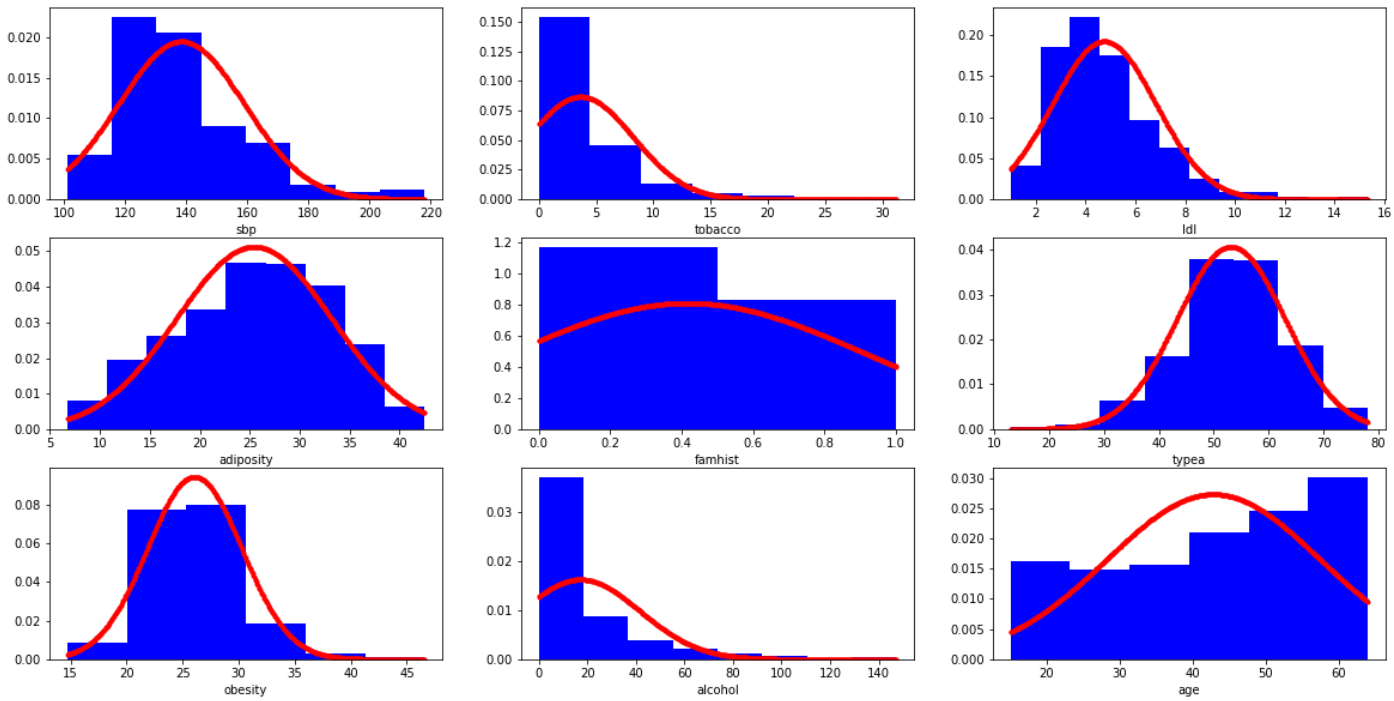
The lower bound is calculated as: Q1 - (1.5*(Q3-Q1))
The upper bound is calculated as: Q3 - (1.5*(Q3-Q1))

The boxplot with potential outliers is shown in Figure 1. Since outliers points do not disturb the bell curve and their values lie within the scope reason, none of them warrant any action.

3.2 Attributes distribution

Looking at the histograms in Figure 2, we could say that sbp, adiposity, typeA and obesity are close to a normal distribution and other attributes i.e tobacco, ldl, famhist, alcohol and age don't seem to have a normal distribution.

Figure 2: Histograms of attributes



But if we use a statistical analysis like the Shapiro-Wilk Test, none of the attributes are strictly gaussian. The p values of the attributes from the Shapiro- Wilk test are as follows: { sbp: 1.25e-14, tobacco: 9.26e-25, adiposity: 7.14e-15,4.24e-05, famhist: 1.87e-30, typea: 0.0086, obesity: 9.22e-10, alcohol: 2.53e-27, age: 4.59e-13}

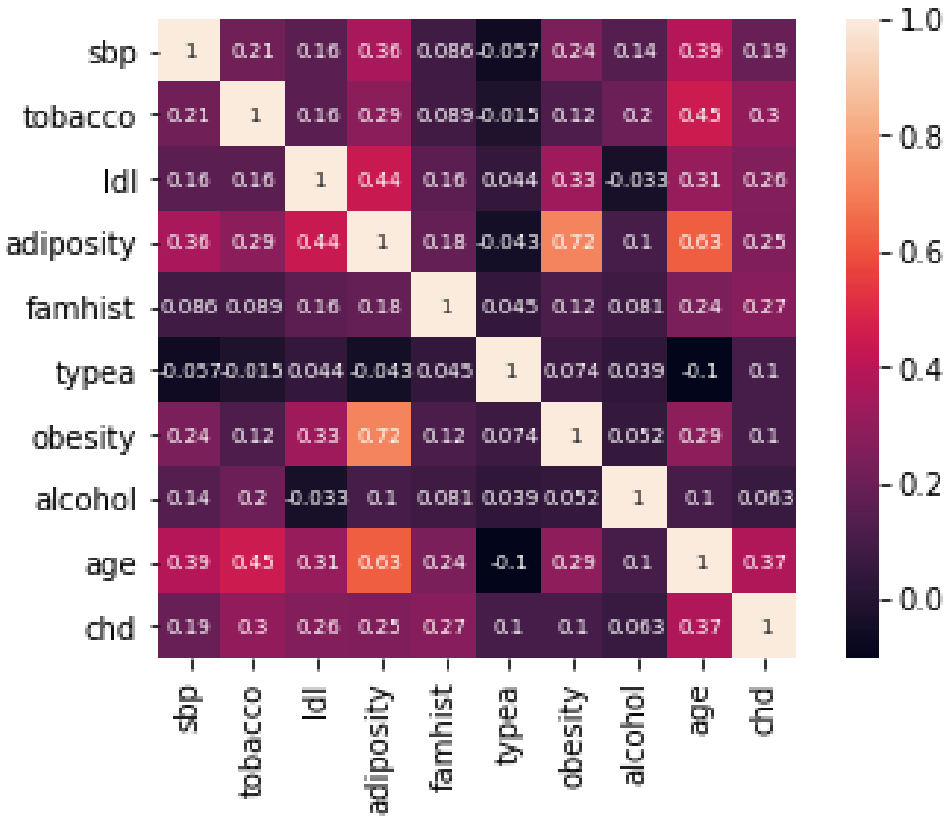
Since the p value for all attributes is less than 0.05, the test won't qualify them as normally distributed.

3.3 Correlation of attributes

In the heat map in Figure 3 we can see the different correlation values of different attributes against each other. Some major findings in this heat map would be as follows:

- Adiposity has a strong correlation with age and obesity.
- Age has a little correlation with almost all attributes except alcohol and typea.
- CHD has the best correlation with age followed by family history, adiposity and ldl where all of the last three mentioned have almost the same value.

Figure 3: Attributes correlation



3.4 Does the primary machine learning modeling aim appear to be feasible based on your visualizations

Since we see high correlation among different attributes especially adisposity, obesity and age, and lower correlation of the attributes with CHD comparatively, the dataset does not seem the most ideal for a classification machine learning model where we would be predicting the probability of CHD with respect to the other attributes.

Also the heatmap tells us that we might need to use a significant number of attributes for prediction of CHD in the classification problem due to its fairly low correlation with the other attributes individually.

Also when we see the correlation coefficient of the typea attribute against other attributes, we see that it has an extremely low correlation with all the other attributes including CHD so it won't be a good idea to do a regression analysis for type a and generally speaking a classification analysis using type a as a dominant factor won't fair well either.

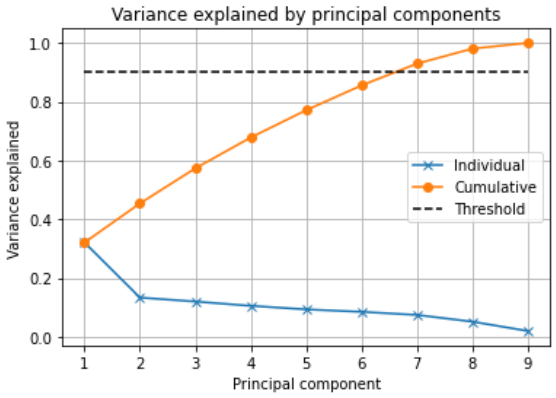
On the other hand, predicting adiposity in the regression model can give us very good results due to its high correlation with a lot of attributes so that seems like a plausible problem which would give us good results.

3.5 PCA analysis

Before carrying out the PCA, first every feature in the data set x was standardized using sklearn.preprocessing python library: x = StandardScaler().fit_transform(x) This function subtracts the mean and then divides the values by standard deviation

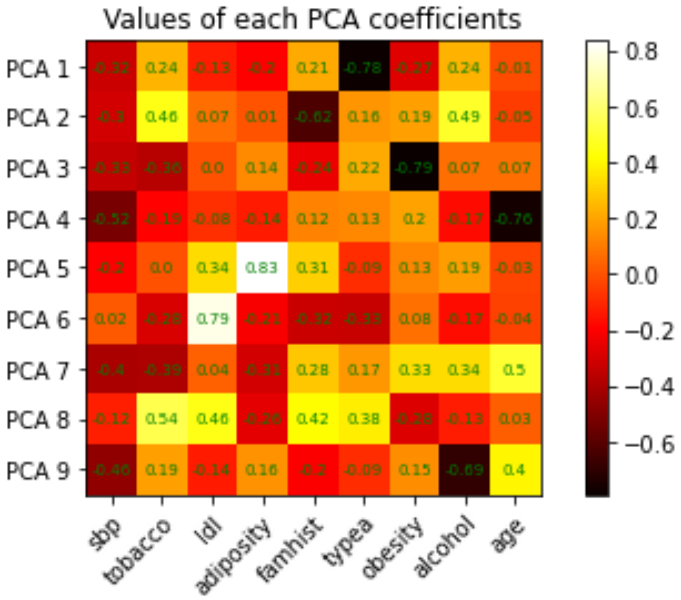
On the graph in the Figure 4 with cumulative variance, it can be seen that in order to keep 90% or more data compliance with the original, minimum 7 first principal components have to be preserved.

Figure 4: Variance of PCA components



Principal directions are 9-dimensional vectors. They are shown on the heatmap in Figure 5 as values of each PCA coefficient. Low absolute value in one direction means that this feature contributes very little to the chosen PCA component. We can see that in the first PCA the most important feature in order to estimate if a person has CHD or not is type A behavior. The minus sign means if somebody has NOT got type A personality, there is also a higher chance he hasn't got CHD.

Figure 5: PCA direction vectors visualization



As mentioned earlier, a minimum of 7 PCA is needed. Plot with only 2 principal components (cumulative variance: 0.45) shows that they are not enough to distinguish if a person has CHD or not. Visualization with 7 dimensions does not make any sense.

4 Conclusion

We learned that even though the histogram plots of all data shows a very close resemblance to a normally distributed curve. A thorough analysis shows that the data points are not exactly normally distributed.

There are very few strong correlations between the features. Even after standardizing the data, only two dimensions of the data could be removed when performing PCA and still keeping the variance above 0.9. So a classification analysis for chd is possible and would give us good results if we use around 7 attributes. Along with that our regression analysis to predict adiposity should be possible based on the correlation heat map. Adiposity has a very good correlation with other attributes, even better than CHD,so a regression analysis should be able to predict the class of adiposity after we have divided it into classes of normal, overweight and obesity etc.

References

[1] LDL optimal values
https://www.mayoclinic.org/tests-procedures/cholesterol-test/about/pac-20384601

[2] Body Adiposity Calculator
https://www.omnicalculator.com/health/bai

[3] South African Heart Disease Data Origin
https://great-northern-diver.github.io/loon.data/reference/SAheart.html#references

[4] BMI explanation
https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[5] "The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani and Jerome Friedman(2009)
https://web.stanford.edu/ hastie/ElemStatLearn/

[6] "Prediction of Coronary Heart Disease by learning from retrospective study" by Srisai Sivakumar
http://srisai85.github.io/CHD/heart_attack.html

[7] "Economic Impact of Advertising Ban on Alcoholic Beverages" Chapter 2 - Alcohol demand/consumption patterns in South Africa
http://www.thedtic.gov.za/wp-content/uploads/nla$_e conomic - impact - of - an - adban.pdf$

[8] Sex, Age, Cardiovascular Risk Factors, and Coronary Heart Disease by Pekka Jousilahti, 1999
https://www.ahajournals.org/doi/full/10.1161/01.CIR.99.9.1165